DOCUMENT RESUME

ED 150 540 CS 003 886

AUTHOR Sakiey, Elizabeth
TITLE A Report of On-Going Research Aimed at Developing Unweighted and Weighted Syllable Lists.
PUB DATE Dec 77
NOTE 10p.; Paper presented at the Annual Meeting of the National Reading Conference (27th, New Orleans, Louisiana, December 1-3, 1977)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Instructional Materials; Language Research; Reading Instruction; Research Methodology; *Structural Analysis; *Syllables; *Word Frequency; *Word Lists

ABSTRACT

Knowing which syllables are most commonly used should aid in linguistic research and in the preparation of curriculum materials, particularly in reading. A research project has been undertaken to develop unweighted and weighted (by the frequency of the words in which they appear) syllable lists. At present, two of the project's three phases are complete. In phase one, the "American Heritage Word Frequency Book" was edited to compile an up-to-date listing of the 5,000 most commonly used words. In phase two, those 5,000 words were coded to indicate the number of letters in each syllable, and this data was merged with the word frequency lists on a computer. The final phase of the project will consist of a computer sort to produce an unweighted syllable list and a summation of syllable frequencies in different words to create the weighted list. (RL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED150540

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Dr. Elizabeth Sakiey
Reading/Speech Correction Dept.
Robinson Academic Building
Glassboro State College
Glassboro, New Jersey 08028
Home: 16 Mae Drive
Yardville, New Jersey 08620

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Elizabeth Sakiey

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

A REPORT OF ON-GOING RESEARCH AIMED AT DEVELOPING UNWEIGHTED AND WEIGHTED SYLLABLE LISTS

Elizabeth Sakiey

At present there is no readily available source to which one can turn to ascertain what the most frequent or basic syllables are. Nonetheless, several attempts have been made in the past to amass quantitative data on syllables.

Frequency Counts of Syllables

Dewey in 1923 listed the 220 commonest phonetic syllables of the 4,400 different syllables found in adult written materials. From elementary school texts of the 1930's, Dolch in 1940 arranged the 100 commonest syllables of polysyllabic words in alphabetical order. Then in 195', Osburn, also utilizing only the polysyllabic words of the elementary grade based Rinsland list (1945), inventoried the initial, medial,

CS003 886

and final syllables.

But not only are these investigations out-of-date, they are inappropriate for most instructional and research needs for other reasons. The Dewey list is not graphemic nor weighted by the words in which the syllables appeared. The Dolch list ignores ranking, is unweighted, and excluded monosyllabic words. Furthermore it is relatively unknown, unused, and not easily obtained. Osburn, too, neglected to weight the syllables by the frequencies of the words in which they appeared. And since nearly half of all words are monosyllabic (Zipf, 1935), logically one-syllable words should be included in a count.

As Dewey (1950) pointed out, "With respect to 'sillables' there exist no previously reliable data on even a small scale" (p. 7). Therefore, there appears to be a need for a graphemic rendition of the most common syllables. One which would be weighted by the frequency of the words in which the syllables occur. Thus a weighted, graphemic syllable frequency count based on an up-to-date source was undertaken.

Procedures

In order to accomplish this a computer tape of the words and their frequencies as they appear in the American Heritage Word Frequency Book (Carroll, Davies, & Richman, 1971) was obtained from the Center for Applied Linguistics in Arlington, Virginia.

Editing a Word Frequency Count

The American Heritage List contained 86,741 different words or "types" drawn from approximately five million words found in curriculum materials ranging from grades three to nine. They are arranged in both Alphabetical and Rank Lists. But since the computer counted as a word

any string of graphic symbols bounded left and right by space, numerals, words with internal punctuation, mathematical symbols, and combinations of these were found among the recognizable words, thereby inflating the list. Moreover, identical words beginning with lower-case and upper-case letters as well as all-capital versions of the same word, such as "boy," "Boy," and "BOY," were treated separately, making the Rank List somewhat misleading. In other words, a word on the Alphabetical List which was recorded three times was consecuently entered in three different positions on the Rank List owing to its three different frequency values.

Thus to determine precisely what the most common 5,000 words were so that a syllable count could be made, the Alphabetical List required editing.

The Edited Alphabetized List. Editing consisted of eliminating 35,000 idiosyncratic words, those with a frecuency of one, of combining like words, and of deleting entries which fell into nonword categories. The eliminated categories were:

1. Abbreviations and acronyms (e.g., Ariz. and AFROTA).
2. Abbreviations with apostrophe of possession (e.g., GE's).
3. Single letters with apostrophe (e.g., A's).
4. Nonword sequences (e.g., bfgc).
5. Words in combination with numbers (e.g., mid-17th).
6. Letters and numbers (e.g., 1A).
7. Roman numerals (e.g., XXX).
8. Dates and other digits (e.g., 1900).
9. All other items not classifiable in other categories (e.g., A-A and E/I).

The printout generated to reflect the editing changes became known as the "Edited Alphabetized List." The data was read onto computer tape also.

The Descending Order Frequency List. The words or types of the Edited Alphabetized List were then ordered by total word frequency using a Fortran sort-merge program. The result was a "Descending Order Frequency List."

Results

After approximately 35,000 idiosyncratic words and 2,000 nonwords were eliminated and 9,000 like-words combined, the 86,741 types of the Alphabetical List of the Word Frequency Book were reduced to 44,174 words. Thus the Alphabetical List was reduced to nearly half its original size.

The differences that one might find in a comparison between the 44,174 word Descending Order Frequency List and the 86,741 word Rank List are due not only to the editing process but also to the fact that the American Heritage Rank List was ordered in terms of the dispersion of the frequencies over 17 subject categories in which the words appeared, such as Reading, English and Grammar, Composition, Literature, Mathematics, Social Studies, Spelling, Science, Music, Art, Home Economics, Shop, Library Fiction, Library Nonfiction, Library References, Magazines, and Religion. These subject categories, however, were of no interest in this investigation.

But despite this difference in approach to ranking, when a comparison was made between the words of the American Heritage List and the shorter Descending Order Frequency List, it was found that there was exact agreement on the ranking of the initial 11 words ("the," "of," "and,"

"a," "to," "in," "is," "you," "that," "it," and "he"). Rankings for the remainder of the first 100 words were similar in all but a few cases.

Significance of the Edited Lists

Even though the Thorndike and Lorge (1944) word frequency list is seriously out-of-date, "it has become the traditional normative source consulted by investigators wishing to manipulate or control for the word frequencies of experimental material" (Shapiro & Gordon, p. 92). That new techniques, such as readability formulas, materials, tests, and experiments are designed using a 1944 word list which actually is an extention of the original 1921 list by Thorndike indicates that what is currently available to researchers is unsuitable. A source of word frequency to serve as the normative source for psychological and educational research is necessary.

The Edited Alphabetized List and the Descending Order Frecuency List can be obtained from the writer upon request.*

Phase Two: Syllabication

The most frequent 5,000 words from the edited Descending Order Frequency List were selected for syllabifying using for reference the American Heritage School Dictionary (1972) and the unabridged American Heritage Dictionary of the English Language (1971). The number 5,000 was selected inasmuch as numerous authorities have stated that is a small number of words that make up most material (Johnson, 1972;

*The Edited Alphabetized List and/or the Descending Order Frecuency List may be obtained at cost plus postage from the author at the Reading/Speech Correction Department, Robinson Academic Building, Glassboro State College, Glassboro, New Jersey 08028.

Jones & Wepman, 1966).

Coding

On computer cards, each of the 5,000 commonest words were punched along with a syllable code number or numbers. The number of letters in each syllable was determined, and that number was placed in the numeric field. For example, a word such as "revolutionary" was coded 312421, so that when scanning from left to right, the computer divided the word as "rev/o/lu/tion/ar/y." This information is organized in Figure 1.

Insert Figure 1 here.

Creation of a Magnetic Disk File

The 5,000 coded syllable cards were then read into the computer and onto a magnetic disk. Then another 5,000 punched cards containing the commonest words and their total word frequencies were read into the computer and merged with the syllabized words on the disk through the use of a Fortran sort-merge program.

Phase Three: Completion of the Project

After the syllables are sorted alphabetically by the computer, a tally will be made of the "tokens" or individual occurrences of each different syllable (type) to produce an unweighted list. Then for each syllable type, the frequencies of the different words in which the syllable appears will be summed to develop a weighted syllable list.

A weighted list is being compiled since it is suspected that many frequently occurring unweighted syllables may not occur in the high

Alpha Field: Columns 1 to 25	Numeric Field: Columns 26 to 31
car	3
player	42
revolutionary	312421

Figure 1. Examples of syllable coding on data cards. For each syllable there is one digit which represents the number of letters in the syllable.

frequency words. Thus an unweighted count alone may be misleading. To date no weighted syllable count exists.

Conclusion

As Durkin (1976) has pointed out "The significance of syllables for phonics is basic: Syllables, not words, are analyzed" (p. 75). Thus knowledge of what the most common syllables are should aid in the preparation of curriculum materials, particularly in reading, and of linguistic research.

REFERENCES

American heritage dictionary of the English language (unabridged). New York: American Heritage, 1971.

American heritage school dictionary. Boston: Houghton Mifflin, 1972.

Carroll, J. B., Davies, P., & Richman, B. American heritage word frequency book. Boston: Houghton Mifflin, 1971.

Dewey, G. Relativ frequency of English speech sounds. Cambridge: Harvard University Press, 1923.

Dewey, G. Relativ frequency of English speech sounds. Cambridge: Harvard University Press, 1950.

Dolch, E. W. Sight syllables versus letter phonics. Elementary School Journal, 1940, 41, 38-42.

Durkin, D. Strategies for identifying words. Boston: Allyn & Bacon, 1976.

Johnson, D. B. Computer frequency control of vocabulary in language learning materials. Instructional Science, Amsterdam, The Netherlands: Elsevier, 1972, 1(1), 121-131. Cited in J. E. DeRocher, M. S. Miron, S. M. Patten, & C. C. Pratt, The counting of words. Syracuse, N. Y.: Syracuse University Research Corporation, 1973.

Jones, L. V., & Wepman, J. M. A spoken word count. Chicago, Ill.: Language Research Associates, 1966.

Osburn, W. J. Teaching spelling by teaching syllables and root words. Elementary School Journal, 1954, 55, 32-41.

Rinsland, H. D. A basic vocabulary of elementary school children. New York: Macmillan, 1945.

Thorndike, E. L., & Lorge, I. The teacher's word book of 30,000 words. New York: Teachers College Press, Columbia University, 1944.

Zipf, G. K. The psycho-biology of language. Boston: Houghton Mifflin, 1935. (Republished: Cambridge, Massachusetts, M. I. T. Press, 1965.)